Metabarcoding and Metagenomics 6: 161–170
DOI 10.3897/mbmg.6.80139

Data Paper

COI amplicon sequence data of environmental DNA collected from the Bronx River Estuary, New York City

Eugenia Naro-Maciel[1], **Melissa R. Ingala**[2], **Irena E. Werner**[3], **Brendan N. Reid**[4], **Allison M. Fitzgerald**[5]

1 *Liberal Studies, New York University, 726 Broadway, New York, New York, 10003, USA*
2 *Sackler Institute for Comparative Genomics, American Museum of Natural History, Central Park West at 79th Street, New York, 10024, USA*
3 *Biology Dept., College of Staten Island, City University of New York, 2800 Victory Boulevard, Staten Island, New York, 10314, USA*
4 *Department of Ecology, Evolution, and Natural Resources, Rutgers University, 14 College Farm Road, New Brunswick, New Jersey 08901, USA*
5 *New Jersey City University, 2039 John F. Kennedy Boulevard, Jersey City, New Jersey, 07305-1597, USA*

Corresponding author: Eugenia Naro-Maciel (enmaciel@nyu.edu)

Academic editor: Florian Leese | **Received** 13 January 2022 | **Accepted** 24 May 2022 | **Published** 10 June 2022

Abstract

In this data paper, we describe environmental DNA (eDNA) cytochrome c oxidase (COI) amplicon sequence data from New York City's Bronx River Estuary. As urban systems continue to expand, describing and monitoring their biodiversity is increasingly important for sustainability. Once polluted and overexploited, New York City's Bronx River Estuary is undergoing revitalization and restoration. To investigate and characterize the area's diversity, we collected and sequenced river sediment and surface water samples from Hunts Point Riverside and Soundview Parks ($n_{total} = 48$; $n_{sediment} = 25$; $n_{water} = 23$). COI analysis using universal primers mlCOIintF and jgHCO2198 detected 27,328 Amplicon Sequence Variants (ASVs) from 7,653,541 sequences, and rarefaction curves reached asymptotes indicating sufficient sampling depth. Of these, eukaryotes represented 9,841ASVs from 3,562,254 sequences. At the study sites over the sampling period, community composition varied by substrate (river sediment versus surface water) and with water temperature, but not pH. The three most common phyla were Bacillariophyta (diatoms), Annelida (segmented worms), and Ochrophyta (e.g. brown and golden algae). Of the eukaryotic ASVs, we identified 614 (6.2%) to species level, including several dinoflagellates linked to Harmful Algal Blooms such as *Heterocapsa* spp., as well as the invasive amphipod *Grandidierella japonica*. The analysis detected common bivalves including blue (*Mytilus edulis*) and ribbed (*Geukensia demissa*) mussels, as well as soft-shell clams (*Mya arenaria*), in addition to Eastern oysters (*Crassostrea virginica*) that are being reintroduced to the area. Fish species undergoing restoration such as river herring (*Alosa pseudoharengus*, *A. aestivalis*) failed to be identified, although relatively common fish including Atlantic silversides (*Menidia menidia*), menhaden (*Brevoortia tyrannus*), striped bass (*Morone saxatilis*), and mummichogs (*Fundulus heteroclitus*) were found. The data highlight the utility of eDNA metabarcoding for analyzing urban estuarine biodiversity and provide a baseline for future work in the area.

Key Words

eDNA, MEGAN, metabarcoding, next-generation sequencing, QIIME2, river sediment, river water, urban ecology

Introduction

Urbanization is increasingly disrupting ecological layouts of cities and their surroundings (Alberti 2008; Douglas and James 2015). Research on urban wildlife can inform strategies to combat related threats such as habitat loss, pollution, and climate change. Further, invasive species and pathogen identification can lead to early action, and conservation planning depends on accurate taxonomic classification.

Despite having one of the world's largest human populations and containing several key habitats such as coastal ecosystems and forests, New York City's

wildlife areas remain insufficiently characterized (Gandy 2003; Sanderson 2009). The Bronx River, which flows through Westchester County and the Bronx, is currently considered 'impaired'. This riparian system is recovering from decades of abuse and still suffering from fecal coliform growth, floating debris, and legacy pollutants such as polychlorinated biphenyls (PCBs), polycyclic aromatic hydrocarbons (PAHs), and metals in the sediments. Combined Sewer Overflow (CSO) drains pump surface run-off and household waste into the river, increasing microplastics and fecal coliforms (NYSDEC 2020). Several local citizen groups host regular cleanups, run reclamation and restoration projects towards targeted species and areas of the river, and educate the public about its resources (e.g., American eels (*Anguilla rostrata*), river herring (*Alosa pseudoharengus*, *A. aestivalis*), and eastern oysters (*Crassostrea virginica*)).

The Estuary Section of the Bronx River Watershed (Fig. 1) contains diverse habitats such as wetlands and streams that face a mix of conservation threats from CSOs and other pollution (NYCParks 2021). Toxins, pathogens, and invasive species occur in urban estuaries, and in the Bronx River several marine and estuarine organisms have established populations. Green crabs (*Carcinus maenas*), Asian shore crabs (*Hemigrapsus sanguineus*), as well as harmful algae that can cause blooms, have all been observed in the river (Fuss and O'Neill 2015). In addition, due to the proximity to roads, housing, and businesses, pathogens that affect humans and marine life (e.g. oyster pathogens *Perkinsus marinus* and *Haplosporidium nelsoni*) can be problematic. To address these issues, habitat assessment, plankton and fish sampling, and water quality and benthic monitoring are in progress (NYCParks 2021). Two key areas of the lower estuary are Hunts Point Riverside Park, a previous garbage dump, and Soundview Park, which borders the estuary and the East River, and is the site of ongoing restoration projects of oysters and salt marshes (Grizzle et al. 2012; Kimmelman 2012; Fitzgerald 2013).

To appropriately characterize and manage such a complex and impacted system, biodiversity inventories and monitoring are key first steps, starting with the correct identification of organisms. Locally in the Bronx and around the world, this has traditionally been achieved through manual surveys requiring organismal capture and/or collection. While providing important information, these methods are potentially labor-intensive and costly, require specific taxonomic expertise, may fail to detect cryptic, microscopic, or elusive taxa, and could provide incorrect or incomplete information. Environmental DNA (eDNA), or DNA sequenced directly from a substrate such as water, sediment, or air, is a flourishing new, non-invasive, rapid, and standardized technology that addresses some of these shortcomings and provides extensive genetic information useful for identifying species through next-generation sequencing (Bik et al. 2012; Bohmann et al. 2014; Taberlet et al. 2018; Deiner et al. 2021).

Biodiversity characterization and monitoring have substantially benefitted from the high quality next-generation bioinformatics pipelines now available to accurately analyze genetic markers with rapidly growing

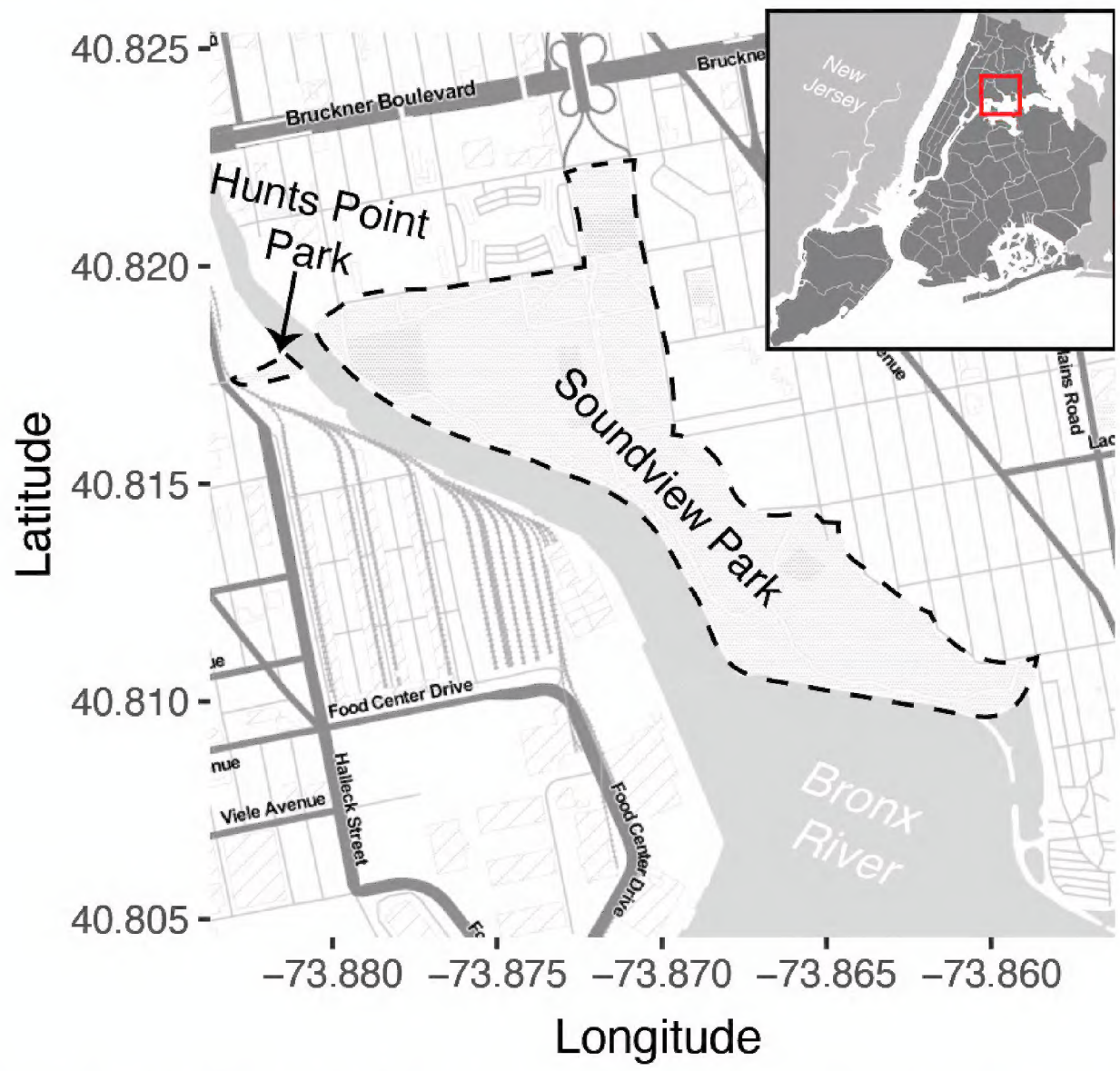

Figure 1. Location of the Hunts Point Riverside and Soundview Park study sites in the Bronx River Estuary (New York City, USA). Samples from each park were collected within 2/10th kilometer. The inset shows the location of the study site (boxed) within the greater New York City metropolitan area. Map data 2019 Google.

reference databases (Taberlet et al. 2018). For instance, our pilot study titled "16S rRNA Amplicon Sequencing of Urban Prokaryotic Communities in the South Bronx River Estuary" revealed the dominant phyla *Proteobacteria*, *Epsilonbacteraeota*, *Cyanobacteria*, *Bacteroidetes*, *Actinobacteria*, and *Acidobacteria*, and found that sediments had higher mean diversity than surface waters (Naro-Maciel et al. 2020). The sequences also contributed to the growing database for the 16S rRNA V4 region, the gold standard employed by the Earth Microbiome Project for prokaryotic metabarcoding (Gilbert et al. 2014). Further, our 18S rRNA gene amplicon (V1–V3 region) Data Paper provided information on an important but less studied 18S region, and successfully amplified a broad range of animals, fungi, and protists (Ingala et al. 2021). We found that community composition varied over time and by substrate (river sediment versus surface water). The sediments were dominated by the phyla Diatomea (diatoms), Annelida, and Nematoda, while the most common phyla in surface waters were Cryptophyceae (algae), Ciliophora (ciliates), Diatomea, and Dinoflagellata. The 18S analysis also detected organisms of management interest such as Eastern oysters (*Crassostrea virginica*) and their pathogens, as well as taxa linked to Harmful Algal Blooms. Although commonly observed species such as soft-shell clams (*Mya arenaria*) and blue mussels (*Mytilus edulis*) were identified, other key common or management-critical taxa such as the fish and invasive species described above were not recovered.

Here we expand our analysis with new COI sequences amplified from the previously analyzed environmental samples (n = 48). In traditional single-species barcoding, COI has been the standard marker for animals due to its conserved priming regions and informatively variable target segment (Hebert et al. 2003). We continued to focus on Amplicon Sequence Variants (ASVs) because the data are reproducible, consistent, and standardized (Callahan et al. 2017), and included distance-based classifications primarily due to incomplete taxonomic databases. Our objectives were to identify organisms, explore biodiversity patterns, and establish a baseline for future work in the Bronx River Estuary.

Methods

Study sites and samples

We sampled benthic sediments and surface waters at Hunts Point (HP, 40.82°N, 73.88°W; $n_{sediment}$ = 9; n_{water} = 8) and Soundview (SVP, 40.81°N, 73.87°W) Parks (Fig. 1), located in Reach 1 of the Bronx River Estuary (NYCParks 2021). At SVP we collected both from a restored oyster reef (SVP-BRO: $n_{sediment}$ = 8; n_{water} = 7) and an area containing wild oysters about one to two tenths of a kilometer distant (SVP-BRC: $n_{sediment}$ = 8; n_{water} = 8). We worked from August 2015 to September 2016, monthly from May–October during low tide as previously described (Fitzgerald 2013; Naro-Maciel et al. 2020; Ingala et al. 2021). We took water pH and temperature measurements using a YSI Pro Plus Probe (YSI, USA) when samples were collected, first at Soundview and later, usually in the same day, at Hunts Point.

DNA metabarcoding

We processed and extracted DNA from these environmental samples within 24 hours as previously described (Naro-Maciel et al. 2020; Ingala et al. 2021). The water samples were filtered with 0.45 µm Whatman Cellulose Nitrate Sterile filters (Cytiva, USA). At the time we did not include extraction blanks or positive controls and worked in a turtle-focused university molecular lab that was not PCR-free (no turtles were detected in this Data Paper). We stringently followed standard decontamination and sterilization procedures in the lab, and later conducted state-of-the-art bioinformatic quality control that removed contaminants and low-quality sequences as discussed below.

A commercial laboratory performed the polymerase chain reaction, clean-up, and sequencing procedures (MRDNA, Molecular Research LP, Shallowater, TX, USA) using previously described industry-standard procedures and controls (Dowd et al. 2008; Naro-Maciel et al. 2020; Ingala et al. 2021). We obtained COI sequences from 48 total samples representing the same river sediment and surface waters samples formerly analyzed for other markers, but in this study amplified with primers mlCOIintF (GGWACWGGWTGAACWGTWTAYCCYCC) (Leray et al. 2013) and jgHCO2198 (TANACYTCNGGRTGNCCRAARAAYCA) (Yu et al. 2012). Polymerase chain reactions were carried out using the Qiagen HotStarTaq Plus Master Mix Kit (Qiagen, USA) with an index on the forward primer, 3 PCR replicates per sample, and standard conditions and controls as reported before (Dowd et al. 2008; Naro-Maciel et al. 2020; Ingala et al. 2021). Following successful 2% agarose gel checks, the samples were pooled in equal proportions and purified with calibrated Ampure XP beads (Agencourt Bioscience, USA). After creating an Illumina amplicon library, an Illumina MiSeq was used to conduct 2 × 300 bp v.3 paired-end sequencing following manufacturer instructions. Samples were sequenced over 3 runs in one initial batch of 33 containing Hunts Point collections and Soundview Park restored oyster reef samples. Later, to add the second Soundview Park site, additional batches of 10 and then the remaining 5 samples from there were processed. All runs produced COI sequences, but due to run-to-run variation the reads produced were shorter in the small last batch. This length variation was dealt with in the bioinformatic processing pipeline as discussed below.

Bioinformatic quality control and analyses

We used the FASTQ Processor to extract indexes and sort forward and reverse reads (MRDNA 2021), and then analyzed raw reads with the QIIME2 v. 2021.4 pipeline of

tools (Bolyen et al. 2019). First, using the DADA2 algorithm (Callahan et al. 2016), reads were joined, dereplicated, chimera-filtered, and then processed as paired-end (Suppl. material 1: Document 1). We ran each sequencing run through DADA2 independently using default parameters (QIIME2 2021), and only after this step were all runs merged into a final, cumulative ASV feature table. We trimmed primers and low-quality base calls from all reads prior to merging with DADA2, and truncated reads to account for declines in quality scores at the sequence ends. DADA2 uses a quality-aware algorithm to identify and correct, if possible, sequencing errors. The software further filters out chimeric sequences and artifacts, leaving behind only joined and dereplicated target sequences (Callahan et al. 2016). For two larger batches a truncation length of 260 bp was used, while for the smallest set of 5 samples a truncation length of 220 bp was used due to shorter overall lengths in this batch. We then aligned the dereplicated sequences using MAFFT (Katoh et al. 2002) and constructed approximate maximum likelihood trees using the FastTree q2-plugin (Price et al. 2010). The average percentage of sequences retained and median of sequences kept per sample are shown in Tables 1, 2.

To assign taxonomic identity to ASVs, a sequence search was conducted against the NCBI database (downloaded 1/27/22) using the blastn algorithm with default parameters in BLAST+ v.2.11.0. (Camacho et al. 2009). BLAST hits were then employed to assign sequences to taxa using the weighted lowest common ancestor, or LCA-assignment algorithm (which identifies the lowest common ancestor in the set of BLAST hits for a given sequence) using MEGAN Community Edition v.6.2.17 (Huson et al. 2016). We used a minimum bitscore of 200 to increase specificity for the LCA analysis. For identifications at the species level we required a minimum 97% match, and we relaxed this to 80% for higher taxonomic levels. Otherwise, default parameters were retained for the LCA analysis. We added phylum-level identifications for any ASVs identified to the order level or lower that did not have these in the NCBI database.

Statistical analyses

We used R v.4.0.0 (R Core Team 2021) as implemented in RStudio v. 1.4.1103 (R Studio Team 2020) for statistical analyses (Suppl. material 1: Document 2). We exported the ASV feature table, taxonomy, rooted phylogeny, and sample metadata to BIOM format and imported these files into R for analysis using the PHYLOSEQ v. 1.32.0 suite of tools (McMurdie and Holmes 2013). First, we identified potential contaminants using the DECONTAM program and filtered them from the ASV feature table (Davis et al. 2018). DECONTAM considers any ASV whose frequency is significantly inversely correlated with sample DNA concentration across all samples as a potential contaminant. We used a conservative threshold of 0.1 to identify significance of contaminants and discarded them from the data set. To assess whether we had sequenced communities deeply enough to detect robust differences in beta diversity, we performed rarefaction analysis using the *rarecurve* function in VEGAN v. 2.5 – 7, and determined that species accumulation curves for all samples had reached asymptotes (Oksanen et al. 2017).

Next, we removed ASVs identified as Archaea (n = 1,185) and Bacteria (n = 12,774) or Domain Unclassified (n = 3,526) from further analysis. We computed sequence abundance-based basic alpha diversity metrics (Observed ASVs, Shannon richness, Faith's phylogenetic diversity, and Pielou's evenness) using a combination of custom functions and commands from the BTOOLS v. 0.0.1 package (Battaglia 2018). We tested for differences in alpha diversity metrics among sites and substrates using the GGPUBR v. 0.4.0 package (Kassambara and Kassambara 2020). We then normalized the data to account for differences in library size among samples by applying the Hellinger transform, which takes the square root of the relative abundance for each taxon and bounds the response between 0 and 1 (Legendre and Gallagher 2001).

We then performed Principal Coordinates (PCoA) ordinations on the abundance-based Bray-Curtis distance matrix and visualized the results by plotting the ordination. 95% confidence ellipses for each site + sample type combination were produced using the stat_ellipse function in ggplot2. To test for turnover in beta diversity among sites and substrates, we performed a PERMANOVA (n_{perm} = 1000) on the Bray-Curtis distance matrix. Because a key assumption of this test is homogeneity of dispersion, we assessed whether our samples met this condition by using the *betadisper* and *permutest* functions in VEGAN. We also tested for the effects of pH, surface water temperature, and year on community composition using a Canonical Correspondence Analysis (CCA) as implemented in VEGAN. Significance was assessed through ANOVA performed on the CCA matrix.

Results and discussion

A total of 48 environmental samples were successfully collected, sequenced, and analyzed for COI (n_{water} = 23; $n_{sediment}$ = 25). Following quality control and contaminant removal, 27,328 ASVs representing Archaea, Bacteria, and Eukarya were recovered from 7,653,541 sequences (Tables 1, 2; Suppl. material 2: Tables S1, S2). Average read depth across samples varied, but in general, high estimates were returned (global minimum: 33,000), and fewer than 1% of ASVs were flagged as contaminants and removed by DECONTAM. Species accumulation curves of each sample reached an asymptote indicative of sufficient sampling depth to detect robust differences in community structure and composition (Suppl. material 2: Fig. S1). Following prokaryote removal, 9,841 ASVs representing algae, animals, fungi, plants, and protists were recovered from 3,562,254 sequences (Suppl. material 2: Table S1). However, data should be interpreted

Table 1. Summary of COI sample data. Sample ID and statistics on the recovery of reads per sample after filtering, denoising, merging, and chimeric sequence removal are displayed, along with the index sequence and sequencing batch. The linker primer sequence for all samples was GGWACWGGWTGAACWGTWTAYCCYCC.

Sample	Index Sequence	input	Filtered	% input passed filter	Denoised	Merged	% of input merged	Non-chimeric	% of input non-chimeric	Batch
S.B.BRC	AATGCAGG	404307	378045	93.5	363757	322683	79.81	305828	75.64	3
S.B.BRO	AATGCTAT	224552	171248	76.26	169030	164525	73.27	160883	71.65	1
S.B.HP	AATGCGAC	264262	197507	74.74	194996	187122	70.81	180612	68.35	1
S.C.BRC	AATGCCGT	446701	419281	93.86	403346	359484	80.48	339110	75.91	3
S.C.BRO	AATTAAGC	230633	172908	74.97	169692	163307	70.81	157334	68.22	1
S.C.HP	AATGTTCG	232082	180166	77.63	176501	168019	72.4	165625	71.36	1
S.D.BRC	AATGCGAC	357594	335024	93.69	320807	284262	79.49	267806	74.89	3
S.D.BRO	AATTATGT	202121	154403	76.39	150533	144088	71.29	142652	70.58	1
S.D.HP	AATTATAA	200934	150589	74.94	147721	141728	70.53	138902	69.13	1
S.E.BRC16	AATCTATT	292305	179623	61.45	161324	140460	48.05	119493	40.88	2
S.E.BRO16	AATTTAGG	261540	203142	77.67	199995	191698	73.3	174901	66.87	1
S.E.HP16	AATTCTCA	225186	170786	75.84	166593	158551	70.41	150573	66.87	1
S.F.BRC16	AATGAGCA	156813	101302	64.6	89909	71922	45.86	64045	40.84	2
S.F.BRO16	AATTTCTA	212778	160538	75.45	158567	153911	72.33	145613	68.43	1
S.F.HP16	ACAAGGCC	239942	181575	75.67	179036	169157	70.5	161451	67.29	1
S.G.BRC16	AATGCAGG	147928	95165	64.33	85196	68675	46.42	60917	41.18	2
S.G.BRO16	ACAATAGA	212958	167753	78.77	165091	159365	74.83	154835	72.71	1
S.G.HP16	ACAATCTG	261294	197228	75.48	193922	185907	71.15	180603	69.12	1
S.H.BRC16	AATGCCGT	176614	111803	63.3	100290	81044	45.89	71905	40.71	2
S.H.BRO16	ACAATTCG	190999	147290	77.12	143306	135326	70.85	133829	70.07	1
S.H.HP16	ACACAAAT	199858	153767	76.94	150265	142697	71.4	139658	69.88	1
S.I.BRC16	AATGCGAC	157544	100475	63.78	90631	73741	46.81	62025	39.37	2
S.I.BRO16	ACACAGCG	180039	138230	76.78	134493	127002	70.54	126083	70.03	1
S.I.HP16	ACACAGGT	251257	191589	76.25	188477	179857	71.58	176210	70.13	1
S.J.HP16	ACACCCAG	296667	220466	74.31	217602	210582	70.98	202212	68.16	1
W.B.BRC	AATCTATT	539622	501610	92.96	493535	470360	87.16	448905	83.19	3
W.B.BRO	AATGAGCA	246358	188743	76.61	185575	175281	71.15	168676	68.47	1
W.B.HP	AATCTATT	219499	173898	79.22	171749	163737	74.6	157321	71.67	1
W.D.BRC	AATGAGCA	497087	462883	93.12	454574	427476	86	403036	81.08	3
W.D.BRO	AATGCAGG	233173	178523	76.56	174353	164570	70.58	160054	68.64	1
W.D.HP	AATGCCGT	163774	123746	75.56	115874	105270	64.28	101538	62	1
W.E.BRC16	AATGCTAT	215382	153878	71.44	134734	124968	58.02	98006	45.5	2
W.E.BRO16	ACACCGGT	234991	184536	78.53	181026	171483	72.97	165766	70.54	1
W.E.HP16	ACACCGAG	180305	137446	76.23	133047	128092	71.04	124416	69	1
W.F.BRC16	AATGTTCG	264274	186824	70.69	164570	153668	58.15	116958	44.26	2
W.F.BRO16	ACAGCGTC	186549	140451	75.29	137606	130972	70.21	127594	68.4	1
W.F.HP16	ACAGCACC	173939	129442	74.42	126420	120337	69.18	116456	66.95	1
W.G.BRC16	AATTAAGC	213600	150028	70.24	131103	121232	56.76	89620	41.96	2
W.G.BRO16	ACAGGGAT	167412	126516	75.57	122513	114505	68.4	109860	65.62	1
W.G.HP16	ACAGTCGT	233719	172803	73.94	168503	159402	68.2	145463	62.24	1
W.H.BRC16	AATTATAA	245212	174443	71.14	152455	141996	57.91	109579	44.69	2
W.H.BRO16	ACAGTTAG	215300	163339	75.87	159885	148340	68.9	142294	66.09	1
W.H.HP16	ACAGTTGC	236955	181379	76.55	178385	168769	71.22	163511	69.01	1
W.I.BRC16	AATTATGT	246099	171387	69.64	149653	137933	56.05	105361	42.81	2
W.I.BRO16	ACATGGCC	229960	178352	77.56	175925	166024	72.2	159860	69.52	1
W.I.HP16	ACATTCTC	228584	177506	77.65	175016	166120	72.67	162118	70.92	1
W.J.HP16	ACATTGAT	210529	162533	77.2	159328	150060	71.28	146483	69.58	1
W.J.SVP16	ACATTGTG	214000	167062	78.07	162840	152186	71.11	147561	68.9	1
TOTALS		11623231						7653541		

with caution given limitations such as a lack of extraction blanks and positive controls, as well as potential lab-related errors in estimating relative abundance (Fonseca 2018; Taberlet et al. 2018). We also note that batch effects in sequencing and analysis can affect interpretation of ASV data (Callahan et al. 2017). Although we did not exhaustively test for sequencing batch effects as this was beyond the scope of this exploratory, data-focused paper, differences among runs could have affected the results and should be accounted for in any future usage of these data. As ASVs retain all variable characteristics of the sequences recovered, including variation in length, counts of ASVs may represent an overestimate of the true number of underlying COI haplotypes, especially since sequences were truncated to different lengths for one batch of sequences.

Table 2. COI sequence and ASV statistics of the Bronx River Estuary. Total or mean values across samples are reported and standard error is shown in parentheses.

Total samples	48
Sample Sites	$HP_{sediment}$ (n = 9) HP_{water} (n = 8) $SVP_{sediment}$ (n = 16) SVP_{water} (n = 15)
Total raw reads	11,623,231
Total reads, passed filter	7,653,541
Raw reads per sample (mean)	242,151 (± 11,768)
Reads per sample, passed filter (mean)	159,449 (± 11,768)
Percent reads passed filter	64.1%
Unique ASVs, pre-filter	27,567
Unique ASVs, contaminants removed	27,328
Total ASVs removed by DECONTAM	239

Variation by substrate, time, and environmental variable

We tested whether there were differences in eukaryotic community composition. There was no significant overall distinction among sites and substrates in phylogenetic diversity (Kruskal-Wallis p = 0.71; Fig. 2A) or observed diversity (Kruskal-Wallis p = 0.7; Suppl. material 2: Fig. S2). There were differences among sites in Shannon diversity (Kruskal-Wallis p = 0.007) and evenness (Kruskal-Wallis p = 0.045), with Hunts Point water showing higher Shannon diversity and evenness than sediments from the same site (Suppl. material 2: Fig. S2). The community turnover (i.e., beta diversity) of eDNA from water samples was significantly different from that of sediment ($r^2 = 0.07$, $P < 0.001$, Fig. 2B). There was no significant differentiation in community composition among sampling years ($r^2 = 0.03$, $P = 0.057$). As regards environmental measurements, at Soundview the average water temperature and pH were 21.1 °C (range 14.5 – 24.2 °C) and 6.9 (range 6.6 – 7.4), respectively. At Hunts Point the average water temperature and pH were 22.6 °C (range 17 – 25.8 °C) and 7.0 (range 6.7 – 7.7). Water temperature had a significant impact on COI community composition ($F_{1,36} = 1.838$, $P = 0.001$; Suppl. material 2: Fig. S3), but there was no significant impact of water pH on river sediment or water profiles ($F_{1,36} = 0.959$, $P = 0.621$).

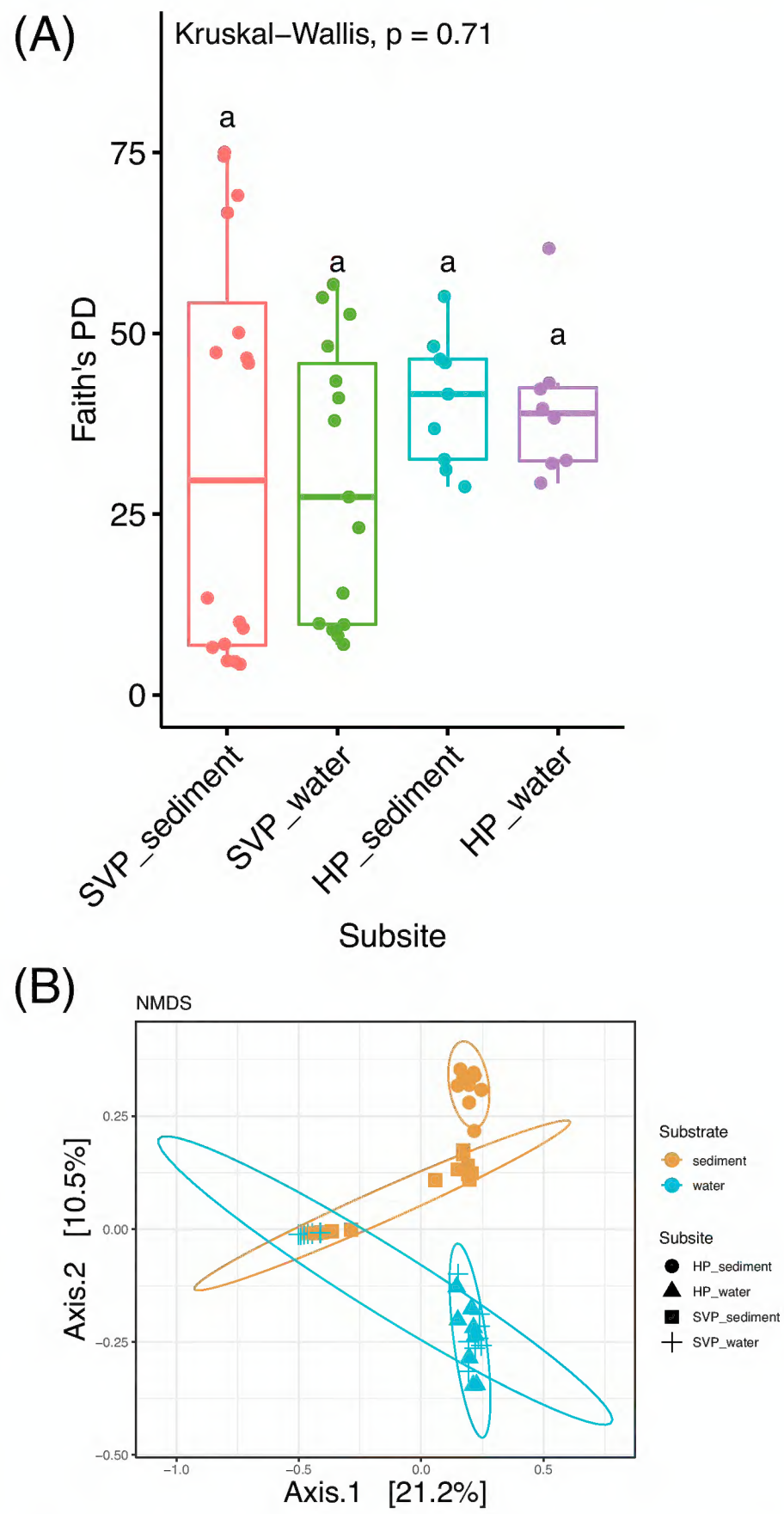

Figure 2. COI diversity comparison between sediment and water samples from Hunts Point (HP) Riverside and Soundview (SVP) Parks. A) Faith's Phylogenetic Diversity. Result of a global Kruskal-Wallace significance test is shown at the top of the plot. Letters indicate no groupings were significantly different from one another based on pairwise significance tests ($p < 0.05$). B) Principal Coordinates Analysis (PCoA) of Bray-Curtis distances. 95% confidence ellipses for each site + sample type combination were produced using the stat_ellipse function in ggplot2.

Comparing eukaryotic eDNA metabarcodes to known Bronx River biodiversity

The analysis detected a variety of common organisms (Fitzgerald 2013; Werner 2016; BRA 2022; iNaturalist 2022), as well as those of management concern, including invasive species (Smithsonian 2022) and potential Harmful Algal Blooms (USNOHAB 2022) (Table 3; Suppl. material 2: Table S1). Of the eukaryotic ASVs, 36.9% were classified to the phylum level or below and 6.2% were classified to species. The gaps in taxonomic resolution are likely linked to a dearth of database information on less studied organisms.

Sequences from the diatom phylum Bacillariophyta were the most commonly detected at most sites and in the dataset overall (Fig. 3; Suppl. material 2: Table S1). Several diatoms were identified including multiple species in the genera *Chaetoceros*, *Lithodesmium*, *Melosira*, *Paralia*, and *Thalassiosira*. For the second most abundant phylum (Annelida) the majority of ASVs mapped to classes Clitellata and Polychaeta. Within Clitellata, the following species were identified: *Amphichaeta sannio*, *Baltidrilus costatus*, *Bothrioneurum vejdovskyanum*, *Limnodrilus hoffmeisteri*, *Monopylephorus rubroniveus*, *Nais elinguis*, *Octolasion cyaneum*, *Paranais litoralis*, *Tubificoides benedii*, *T. brownie*, *T. fraseri*, and *T. parapectinatus*.

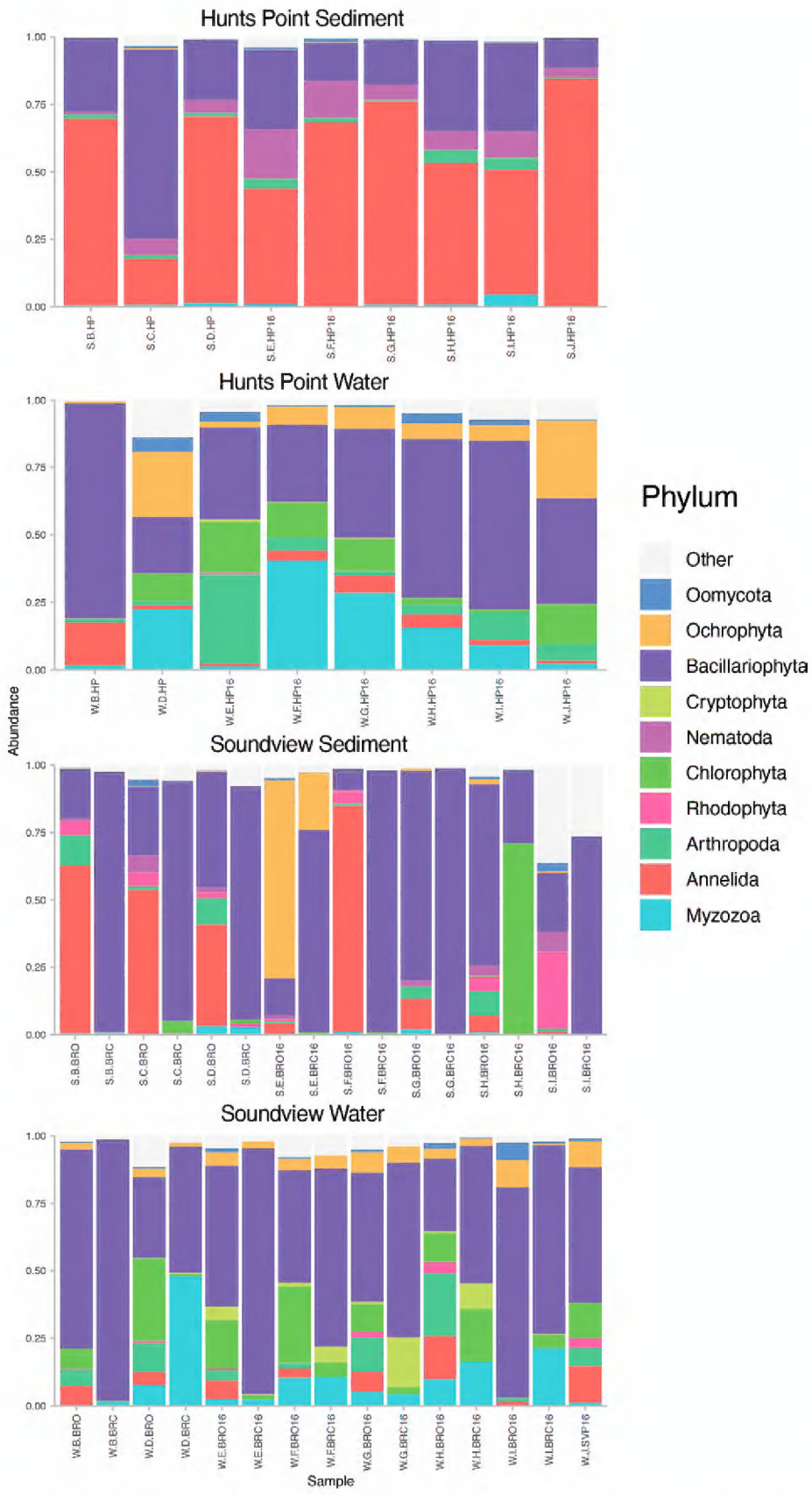

Figure 3. Community profiles of eukaryotic COI Amplicon Sequence Variants (ASVs) in sediment and water samples from Hunts Point Riverside and Soundview Parks. Shown in temporal order of collection at the level of phylum; bar heights indicate relative abundance of sequences from each taxon.

Worms of Class Polychaeta identified to species were: *Amphitrite ornate*, *Capitella teleta* (common in the area), *Glycera americana* (American bloodworm), *Glycinde multidens*, *Hypereteone heteropoda*, *Parasabella microphthalma*, *Polydora cornuta*, and *Streblospio benedicti* (the common Ram's horn worm).

Further, several key organisms being restored and monitored in the Bronx River, as well as commonly observed species, were detected (Table 3; Suppl. material 2: Table S1). For example, sequences exhibiting a perfect match to American eels (*Anguilla rostrata*) were found, although these data also matched sequences assigned to the family level in NCBI. Other fish species whose presence is associated with healthy tidal areas, including mummichogs (*Fundulus heteroclitus*), Atlantic silversides (*Menidia*

Table 3. Eukaryotic species of special interest detected by COI from the Bronx River Estuary. C = Commonly observed; M = of Management Concern.

Class and Genus species	Common name	TYPE	COI
Actinopterygii			
Alosa pseudoharengus	River Herring/Alewife	M	
Alosa estivalis	River Herring	M	
Ameiurus nebulosus	Brown bullhead	C	
Ameiurus spp.	Bullhead catfish	C	✓
Anguilla rostrata	American eel	M	✓
Brevoortia tyrannus	Menhaden	C	✓
Fundulus heteroclitus	Mummichog	C	✓
Fundulus majalis	Striped Mummichog (or killifish)	C	
Gobiesox strumosus	Skillet fish	C	
Lepomis spp.	Sunfish	C	
Menidia menidia	Atlantic silverside	C	✓
Morone americana	White perch	C	✓
Morone saxatilis	Striped bass	C, M	✓
Perca flavescens	Yellow perch	C	✓
Ascidiacea			
Botryllus schlosseri	Golden star tunicate	C	✓
Molgula spp.	Sea grape	C	
Perophora sagamiensis	Sea squirt	C	
Aves			
Branta canadensis	Canada goose	C	
Egretta spp.	Egrets, Herons	C	
Larus spp.	Gulls	C	✓
Bivalvia			
Crassostrea virginica	Eastern oyster	M	✓
Euglesa casertana	Pea Clam		
Geukensia demissa	Ribbed mussel	C	✓
Macoma petalum	Atlantic Macoma		✓
Mercenaria mercenaria	Hard or chowder clam	C	
Mulinia lateralis	Dwarf surf clam	C	✓
Mya arenaria	Soft-shell clam	C	✓
Mytilus edulis	Blue mussel	C	✓
Nucula proxima	Atlantic nut clam		✓
Petricolaria pholadiformis	False angelwing	C	✓
Demospongiae			
Cliona spp.	Boring sponge	C	
Halichondria panicea	Breadcrumb sponge	C	✓
Dinophyceae			
Alexandrium spp.	HAB (potential)	M	
Amphidinium carterae	HAB (potential)	M	
Dinophysis sacculus	HAB (potential)	M	
Gymnodinium spp.	HAB (potential)	M	
Gyrodinium spp.	HAB (potential)	M	✓
Heterocapsa rotundata	HAB (potential)	M	✓
Heterocapsa triquetra	HAB (potential)	M	✓
Heterocapsa spp.	HAB (potential)	M	✓
Karlodinium sp. RS-24	HAB (potential)	M	✓
Margalefidinium polykrikoides	HAB (potential)	M	✓
Gastropoda			
Corambe obscura	Obscure Corambe		✓
Crepidula fornicata	Common slipper snail	C	
Ercolania fuscata	Sea Slug		✓
Tritia obsoleta (syn *Ilyanassa obsoleta*)	Eastern mudsnail	C	✓
Urosalpinx cinerea	Oyster drill	C	
Malacostraca			
Callinectes sapidus	Blue crab	C, M	
Carcinus maenas	Green crab	C, M	

Class and Genus species	Common name	TYPE	COI
Dyspanopeus sayi	Mud crab	C	
Gammarus oceanicus	Scud amphipod	C	
Grandidierella japonica	Invasive amphipod	M	✓
Hemigrapsus sanguineus	Asian shore crab	C, M	
Microdeutopus gryllotalpa	Slender tube maker	C	
Pagurus longicarpus	Long-clawed hermit crab	C	
Palaemonetes pugio	Common shore shrimp	C	
Panopeus herbstii	Black fingered mud crab	C	
Rhithropanopeus harrisii	White fingered mud crab	C	
Mammalia			
Homo sapiens	Human	C	✓
Ondatra zibethicus	Muskrat	C	
Rattus norvegicus	Brown rat	C	✓
Merostomata			
Limulus polyphemus	Horsheshoe crab	C, M	✓
Polychaeta			
Alitta succinea (syn *Nereis succinea*)	Clam worm	C	
Amphitrite ornata	Ornate worm		✓
Capitella teleta	Thread worm	C	✓
Glycera americana	Blood worm		✓
Lycastopsis pontica	Spring worm	C	
Platynereis dumerilii	Dumeril's clam worm	C	
Streblospio benedicti	Ram's horn worm	C	✓

menidia), and menhaden (*Brevoortia tyrannus*), were detected. River herring (*Alosa pseudoharengus* and *A. aestivalis*), however, were not identified. Arthropod ASVs included a non-native belostomatid water bug (*Appasus major*), the invasive malacostracan *Grandidierella japonica*, and *Limulus polyphemus*, the Atlantic horseshoe crab. Various bivalves were present, including the eastern oyster, *Crassostrea virginica*, which has been the focus of targeted restoration efforts in New York City waterways, in addition to commonly observed blue (*Mytilus edulis*) and ribbed (*Geukensia demissa*) mussels, and soft-shell clams (*Mya arenaria*). Dinoflagellate taxa potentially linked to harmful algal blooms were also recovered including in the genus *Heterocapsa*. In conclusion, this COI Data Paper complements our prior 16S and 18S pilot work (Naro-Maciel et al. 2020; Ingala et al. 2021), and provides a baseline for future metabarcoding efforts to characterize urban estuarine biodiversity in the Bronx River, with applications for other areas.

Data availability

All amplicon gene sequences from this study are posted on the NCBI Sequence Read Archive (SRA) under BioProject PRJNA606795. DNA extracts are stored at the American Museum of Natural History.

Conflicts of interest

The authors declare no competing interests.

Acknowledgements

We are grateful to the New York University (NYU) Research Challenge Fund and NYU Liberal Studies New Faculty Scholarship and Creative Production Awards (to ENM), and to private donors through Experiment.com (to IW) for funding the research. Site access was provided by NY/NJ Baykeeper and the New York City Department of Parks and Recreation (Natural Resources Group). Our special thanks to Michael Tessler for initial guidance, as well as student assistants Christian Bojorquez, NaVonna Turner, Sean Thomas, Jennifer Servis, Patrick Shea, Vanessa Van Deusen, and Seth Wollney. We are very thankful for two anonymous reviewers, Reviewer Lise Klunder, and our Subject Editor Florian Leese, whose helpful comments improved the manuscript.

References

Alberti M (2008) Advances in urban ecology. Springer, Seattle, 366 pp. https://doi.org/10.1007/978-0-387-75510-6

Battaglia T (2018) btools: A suite of R function for all types of microbial diversity analyses. R package version 0.0.1.

Bik HM, Porazinska DL, Creer S, Caporaso JG, Knight R, Thomas WK (2012) Sequencing our way towards understanding global eukaryotic biodiversity. Trends in Ecology & Evolution 27(4): 233–243. https://doi.org/10.1016/j.tree.2011.11.010

Bohmann K, Evans A, Gilbert MTP, Carvalho GR, Creer S, Knapp M, Yu DW, de Bruyn M (2014) Environmental DNA for wildlife biology and biodiversity monitoring. Trends in Ecology & Evolution 29(6): 358–367. https://doi.org/10.1016/j.tree.2014.04.003

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS II, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37(8): 852–857. https://doi.org/10.1038/s41587-019-0209-9

BRA (2022) What Lives in the River? Bronx River Alliance. https://bronxriver.org/restoration-access/about-the-river/what-lives-in-the-river [January 7, 2022]

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Amy Jo A (2016) DADA2: High resolution sample inference from Illumina amplicon data. Nature Methods 13(7): 48–56. https://doi.org/10.1038/nmeth.3869

Callahan BJ, McMurdie PJ, Holmes SP (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal 11(12): 2639–2643. https://doi.org/10.1038/ismej.2017.119

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST +: Architecture and applications. BMC Bioinformatics 10(1): e421. https://doi.org/10.1186/1471-2105-10-421

Davis NM, Proctor D, Holmes SP, Relman DA, Callahan BJ (2018) Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome 6(1): 221499–221499. https://doi.org/10.1186/s40168-018-0605-2

Deiner K, Yamanaka H, Bernatchez L (2021) The future of biodiversity monitoring and conservation utilizing environmental DNA. Environmental DNA 3(1): 3–7. https://doi.org/10.1002/edn3.178

Douglas I, James P (2015) Urban Ecology: An Introduction. Routledge, Taylor and Francis Group, New York, New York, 500 pp.

Dowd SE, Sun Y, Wolcott RD, Domingo A, Carroll JA (2008) Bacterial Tag-Encoded FLX Amplicon Pyrosequencing (bTEFAP) for Microbiome Studies: Bacterial Diversity in the Ileum of Newly Weaned Salmonella -Infected Pigs. Foodborne Pathogens and Disease 5(4): 459–472. https://doi.org/10.1089/fpd.2008.0107

Fitzgerald AM (2013) The effects of chronic habitat degradation on the physiology and metal accumulation of eastern oysters (*Crassostrea virginica*) in the Hudson Raritan Estuary. PhD Thesis. Graduate Center, The City University of New York.

Fonseca VG (2018) Pitfalls in relative abundance estimation using eDNA metabarcoding. Molecular Ecology Resources 18(5): 923–926. https://doi.org/10.1111/1755-0998.12902

Fuss, O'Neill (2015) Citizen Science on the Bronx River: An Analysis of Water Quality Data. Bronx, New York, 57 pp.

Gandy M (2003) Concrete and Clay: Reworking Nature in New York City. MIT Press, Boston, Massachusetts, 358 pp. https://doi.org/10.7551/mitpress/2083.001.0001

Gilbert JA, Jansson JK, Knight R (2014) The Earth Microbiome project: Successes and aspirations. BMC Biology 12(1): e69. https://doi.org/10.1186/s12915-014-0069-1

Grizzle R, Ward K, Lodge J, Suszkowski D, Mosher-Smith K, Kalchmayr K, Malinowski P (2012) Oyster Restoration Research Project (ORRP) Technical Report. New York, New York.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London. Series B, Biological Sciences 270(1512): 313–321. https://doi.org/10.1098/rspb.2002.2218

Huson DH, Beier S, Flade I, Górska A, El-hadidi M (2016) MEGAN Community Edition – Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. Computational Biology 12: e1004957. https://doi.org/10.1371/journal.pcbi.1004957

iNaturalist (2022) Bronx River Parkway Check List. https://www.inaturalist.org/check_lists/998383-Bronx-River-Parkway-Check-List?page=3 [January 7, 2022]

Ingala MR, Werner IE, Fitzgerald AM, Naro-Maciel E (2021) 18S rRNA amplicon sequence data (V1–V3) of the Bronx river estuary, New York. Metabarcoding and Metagenomics 5: e69691. https://doi.org/10.3897/mbmg.5.69691

Kassambara A, Kassambara MA (2020) Package 'ggpubr.' R package version 0.4.0.

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30(14): 3059–3066. https://doi.org/10.1093/nar/gkf436

Kimmelman M (2012) Bronx River Now Flows by Parks. The New York Times.

Legendre P, Gallagher ED (2001) Ecologically meaningful transformations for ordination of species data. Oecologia 129(2): 271–280. https://doi.org/10.1007/s004420100716

Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ (2013) A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. Frontiers in Zoology 10(1): e34. https://doi.org/10.1186/1742-9994-10-34

McMurdie PJ, Holmes S (2013) Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE 8(4): e61217. https://doi.org/10.1371/journal.pone.0061217

MRDNA (2021) FASTQ Processor. http://www.Mrdnafreesoftware.com [July 8, 2021]

Naro-Maciel E, Ingala MR, Werner IE, Fitzgerald AM (2020) 16S rRNA Amplicon Sequencing of Urban Prokaryotic Communities in the South Bronx River Estuary. Microbiology Resource Announcements 9(22): e00182-20. https://doi.org/10.1128/MRA.00182-20

NYCParks (2021) Reach 1: Downstream of Weir – Estuary Section: Wetlands of the Bronx River Watershed. https://www.nycgovparks.org/greening/natural-resources-group/ronx-river-wetlands/estuary-section/reach-1 [June 23, 2021]

NYSDEC (2020) NYS Section 303(d) List of Impaired/TMDL Waters – NYS Dept. of Environmental Conservation. https://www.dec.ny.gov/chemical/31290.html [June 9, 2021]

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Peter Solymos M, Stevens HH, Szoecs E, Wagner H (2017) vegan. Community Ecology Package 2: 4–5.

Price MN, Dehal PS, Arkin AP (2010) FastTree 2 – Approximately maximum-likelihood trees for large alignments. PLoS ONE 5(3): e9490. https://doi.org/10.1371/journal.pone.0009490

QIIME2 (2021) QIIME2 denoise-single: Denoise and dereplicate single-end sequences. https://docs.qiime2.org/2021.4/plugins/available/dada2/denoise-single/ [July 19, 2021]

R Core Team R (2021) A language and environment for statistical computing. https://www.r-project.org/ [October 21, 2021]

R Studio Team (2020) RStudio: Integrated Development for R. http://www.rstudio.com/ [October 21, 2021]

Sanderson EW (2009) Mannahatta: A Natural History of New York City. Harry N. Abrams, New York, New York, 352 pp.

Smithsonian (2022) Nemesis. https://invasions.si.edu/nemesis/ [January 12, 2022]

Taberlet P, Bonin A, Zinger L, Coissac E (2018) Environmental DNA: For Biodiversity Research and Monitoring. Oxford University Press, London, 253 pp. https://doi.org/10.1093/oso/9780198767220.001.0001

USNOHAB (2022) Harmful Algal Bloom Species by Name. US National Office for Harmful Algal Blooms. https://hab.whoi.edu/species/species-by-name/ [January 12, 2022]

Werner I (2016) Assessing urban oyster restoration through classical and next-generation approaches. Master's Thesis. College of Staten Island, The City University of New York.

Yu DW, Ji Y, Emerson BC, Wang X, Ye C, Yang C, Ding Z (2012) Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods in Ecology and Evolution 3(4): 613–623. https://doi.org/10.1111/j.2041-210X.2012.00198.x

Supplementary material 1
Supplementary Data Files 1, 2
Author: Brendan Reid, Melissa Ingala
Data type: QIIME AND R Scripts
Explanation note: Scripts used for metabarcoding analysis. **Document 1:** QIIME2 workflow; **Document 2:** R script.

Link: https://doi.org/10.3897/mbmg.6.80139.suppl1

Supplementary material 2
Tables S1, S2, Figures S1–S3
Author: Brendan Reid, Melissa Ingala
Data type: Figures and tables
Explanation note: **Fig. S1.** Sample-based species accumulation curves of COI Amplicon Sequence Variant (ASV) diversity by substrate type (sediment, water). Calculated using the VEGAN 2.4-3 package. **Fig. S2.** Eukaryotic alpha diversity comparison between sites and substrate types. Measured by COI for Observed ASVs, Shannon richness, and Pielou's evenness. Results of a global Kruskal-Wallace significance test are shown at the top of each plot. Letters indicate groupings that were significantly different from one another based on pairwise significance tests ($p < 0.05$). **Fig. S3.** Canonical Correspondence Analysis indicating the influence of water temperature and pH on eukaryotic community composition inferred by COI. Study sites (Hunts Point (HP) and Soundview (SVP) Parks) and substrates (sediment, water) are shown as different shapes, and arrow lengths indicate the strength and direction of the influence. **Table S1.** Taxonomic Assignment including COI ASV identification to Domain, Kingdom, Phylum, Class, Order, Family, Genus, and/or Species. ASVs identified as putative contaminants are included at the end of the table. **Table S2.** Taxonomic Assignment Totals of COI ASVs (number and percentage) identified to Domain, Kingdom, Phylum, Class, Order, Family, Genus, and/or Species.

Link: https://doi.org/10.3897/mbmg.6.80139.suppl2